CNN-ELM 混合短文本分类模型 *

韩众和 [1]，夏战国 [1]，杨 婷 [2]

(1. 中国矿业大学 计算机科学与技术学院，江苏 徐州 221000; 2. 中科院电子学研究所苏州研究院，江苏 苏州 215000)

摘 要： 针对目前自然语言处理研究中，使用卷积神经网络（CNN）进行短文本分类任务时可以结合不同神经网络结构与分类算法以提高分类性能的问题，提出了一种结合卷积神经网络与极速学习机的 CNN-ELM 混合短文本分类模型。使用词向量训练构成文本矩阵作为输入数据，然后使用卷积神经网络提取特征并使用 Highway 网络进行特征优化，最后使用误差最小化极速学习机(EM-ELM)作为分类器完成短文本分类任务。与其他模型相比，该混合模型能够提取更具代表性的特征并能快速准确地输出分类结果。在多种英文数据集上的实验结果表明提出的 CNN-ELM 混合短文本分类模型比传统机器学习模型与深度学习模型更适合完成短文本分类任务。

关键词： 文本分类；卷积神经网络；极速学习机

中图分类号： TP391.1 doi: 10.3969/j.issn.1001-3695.2017.09.0930

Hybrid CNN-ELM model for short text classification

Han Zhonghe[1], Xia Zhanguo[1], Yang Ting[2]

(1. *College of Computer Science & Technology*, *China University of Mining & Technology*, *Xuzhou Jiangsu* 221000, *China*; 2. *Institute of Electronics*, *Chinese Academy of Science*, *Suzhou Jiangsu* 215000, *China*)

Abstract: In current Natural Language Processing research, people can combine different neural network structure and classification algorithm when using Convolution Neural Network (CNN) to conduct text classification tasks so as to improve the classification performance. Thus, this paper proposed a hybrid CNN-ELM model for short text classification. Firstly, the model used word vectors to represent sentence as the input data. Secondly, it extracted features through CNN and completed features optimization with Highway network. Finally, it used error minimization extreme learning machine (EM-ELM) as a classifier to complete text classification task. Compared with other models, the proposed model could extract more representative features and output classification results more quickly and accurately. According to the experimental results in various English data sets, the proposed model is more suitable for short text classification tasks than traditional machine learning models and deep learning models.

Key Words: text classification; convolutional neural networks; extreme learning machine

0 引言

深度学习近年来在自然语言处理（NLP）的应用越来越广泛，短文本分类更是其中重要的一部分。短文本分类是指对有价值的短文本信息进行分类处理，其在目前的信息社会中有着非常重要的意义。在短文本分类中最关键的问题是文本特征的提取，传统的特征提取方法诸如 MI[1]、pLSA[2]、LDA[3]等会忽略文本中的上下文关联信息从而不能准确获取词汇的语义。

近年来，深度学习在图像识别与手写识别的惊人表现有目共睹。但若要使用深度学习实现文本处理就需要将文本进行数字化的表示。词嵌套（word embedding）是目前最有效的保留词汇语法、语义信息的词向量转换方式。这种使用训练过程的算法将词汇的相似性表达为向量空间的相似度，极大地保留了词汇的语义与语境信息[4,5]。借助于词嵌套的帮助，对文本使用深度学习以提取丰富的特征就成为了可能。

卷积神经网络（CNN）作为深度学习中最具代表性的结构之一，其在文本处理中的应用已相当广泛。在这些 CNN 结构当中，Kim 于 2014 年提出的 CNN 模型[6]很好的证明了词向量应用在简单 CNN 结构上所能对文本分类产生的巨大影响，这也使得该模型成为了 CNN 应用在自然语言处理中最具代表性的模型之一。在此之后，Mandelbaum 等人在 TensorFlow 上实现了 Kim 的模型并进行了相关改进[7],进一步提高了 CNN 模型

基金项目： 国家自然科学基金资助项目（61572506）

作者简介： 韩众和（1993-），男，河南郑州人，硕士研究生，主要研究方向为深度学习、自然语言处理（hanzhonghe317@163.com）；夏战国（1974-），男，副教授，硕导，博士，主要研究方向为机器学习、数据挖掘；杨婷（1992-），女，硕士研究生，主要研究方向为自然语言处理.

在多种英文分类数据集上的分类精确度。在国内的研究中，陈钊等人[8]将情感词典识别构成的二值特征与 CNN 提取的特征相结合，这种添加外部辅助特征的方法显著提高了 CNN 模型的情感分析能力；刘龙飞等人[9]证明了使用字级别特征进行 CNN 情感分析比词级别更有效；彤博辉等人[10]将不同通道的词向量相结合作为 CNN 的输入来完成实体关系抽取工作，该模型的宏平均 F1 值比 CNN、RNN 模型更高。这些方法大多集中在 CNN 结构的输入与特征表达的改进上，除此之外，将 CNN 特征提取模型与不同的分类原理相结合亦能够有效的提升模型的性能。

目前，将 CNN 与不同的分类原理相结合的最常见方式是将 CNN 与 SVM 相结合，这种方法已经被应用在情感分析与人脸识别中并获得了比传统 CNN 分类模型更好的结果[11,12]。然而，在实验中使用交叉验证法（Cross Validation）划分训练与测试集时,SVM 会产生较大的时间损耗来确定其自身的参数，同时其性能依旧有提升的空间。

黄广斌提出的极速学习机（ELM）[13]是一种强大的机器学习模型，它是一种可以随机选择隐层节点数并计算输出权重的单隐层前馈神经网络（SLFNs），它的特点是泛化能力强且拥有非常快的学习速度，有研究表明 ELM 分类器比 SVM 分类器更优秀[14]。近年来，将极速学习机与卷积特征相结合的方法在个别领域中得到了实现[15]，但还未有人在文本处理中进行相关研究。极速学习机虽然有着优异的泛化能力，但获取极速学习机的最优结果大多是通过人工统计来实现的。作为极速学习机的改进之一，误差最小化极速学习机（EM-ELM）[16]的提出使极速学习机具备了自动计算最优解并持续更新网络输出权重的能力。因此，将卷积神经网络与误差最小化极速学习机相结合以提高短文本分类性能就成为了可能。

受 LSTM 网络的启发，Srivastava 等人[17]于 2015 年提出了 Highway 网络，通过对深度神经网络提取的特征进行类似门结构的特征优化，该网络能够有效解决深度学习中多层网络训练时难以有效收敛的问题。在具体应用上，R.K. Srivastava 等人将这种新的网络结构分别与 CNN 及全连接网络相结合并以此进行了相关实验[18]，在图像识别数据集上（CIFAR,MNIST 等）的实验结果表明，结合了 Highway 网络的深度神经网络获得了更高的精确度。

除图像处理之外，Kim 将 Highway 网络应用在自然语言处理中，他提出的模型将 CNN 与 Highway 网络、LSTM 相结合并使用字符级输入完成多语种语言分析任务[19]。同时他也分析了 Highway 网络在该模型中的重要性，在 PTB 数据集上的实验表明 Highway 网络能够对 CNN 提取出的特征进行优化且该模型中 Highway 网络为 2 层时性能最优。在语音识别的应用上，Wei-Ning Hsu 等人[20]在使用 CNN、LSTM、DNN 三种模型构成 CLDNN 模型的基础上，在 LSTM 中加入 Highway 网络实现加深网络层数与优化网络特征的功能。这种新的模型在中文广播语音数据集上获得了超越以往任何模型的最优性能。

综上所述，本文提出一种 CNN-ELM 混合短文本分类模型，和 CNN 与 ELM 在图像分类中的结合[15]不同，本文模型使用一维卷积方式获取特征向量，再结合多层 Highway 网络构成深度网络。同时，对 ELM 的改进也不仅局限在优化参数的初始化方法上[15]，因为这种改进方式依旧需要大量的隐层节点数测试来获取最优结果。本文模型通过使用 EM-ELM 实现直接输出最优结果的功能，该方法能够避免大量测试所需的时间损耗。

实验结果表明，本文模型在多种英文短文本分类数据集上的性能比传统机器学习分类模型和传统卷积神经网络模型更优秀。本文创新点为：将卷积神经网络与极速学习机原理相结合，提高了模型的泛化能力，有效提高了分类效果；在混合模型中加入 Highway 网络层进行特征优化，进一步提高了模型的性能，同时也研究了不同 Highway 网络层数对本文混合模型分类性能的影响。

1　数据预处理与文本的词向量表示

1.1　数据预处理

由于本文使用英文短文本数据集，这些短文本中的语言表达具有网络语言相对特殊的属性，所以这使得短文本分类任务面临诸多挑战。本文在进行数据预处理时，普遍采用清洗文本的手段来将标点符号与不相关的符号剔除，从而减轻词向量转换与分类工作的工作量。考虑到短文本数据集中存在较多的网页地址、符号、符号表情等复杂属性，本文使用正则表达式识别网址标记（http）、@符号、话题符号(#)与简单的符号表情，并将非英文文字转换为特征标记。最后，本文将所有未识别的符号与文字归为超出词典的情况，并在词向量训练时将其转换为具有随机值的词向量。

1.2　文本的词向量表示

词向量的定义有很多种，其中最广为认可的是词向量是一个词的数字化表示，这通常是以向量的形式呈现出来的。更准确的来说，词向量是将一个词所代表的语义与对该词通过非监督训练方法得到的向量关联起来的一种技术。与图像本身就是丰富的、高维度的向量不同，文本语言不能直接作为数据挖掘算法的输入数据来进行更深层次分析，所以将其进行离散化的表示就有着极其重要的意义。

最初，人们使用 One-hot 词向量来表示词汇，这种方法使用一个向量来表示一个词，向量的长度为词典的大小，向量的分量只有一个 1，其他全为 0，其中分量 1 的位置对应该词在词典中的位置。这种方法虽简单易行却也有着明显的缺点：a）容易受维数灾难的困扰，对短文本数据采用这种方法构建词向量往往会造成词向量非常稀疏,尤其是将其用于深度学习中的一些算法时；b）不能很好地刻画词与词之间的相似性,即“词汇鸿沟”现象：任意两个词之间都是孤立的，单从这两个向量并不能看出这两个词是否有联系。

为了弥补 One-hot 词向量的不足，Hinton 提出了一种叫做 word embedding 的词向量表示方法，这种方法的主要思想是将

词分布式地映射到低维空间中，从而解决了向量稀疏性的问题。此外，该低维空间中词向量之间的位置关系可以很好地反映它们在语义层面上的联系，使其非常适合作为文本的高层抽象特征。

目前，大部分人使用 Google 于 2013 年发布的用于训练词向量的软件工具 Word2Vec[1]。它根据给定的语料库，通过优化后的训练模型快速有效的将一个词语表达成向量形式，该工具的核心架构包括 Continuous bag-of-word (CBOW)和 Skip-gram。就原理上来讲，CBOW 从上下文语义中预测目标词，而 Skip-gram 使用目标词汇来预测句子文本的上下文语义。本文使用 Skip-gram 模型来训练词向量，因为在大容量的数据集上它的性能比 CBOW 更好。该词向量训练模型结构如图 1 所示：

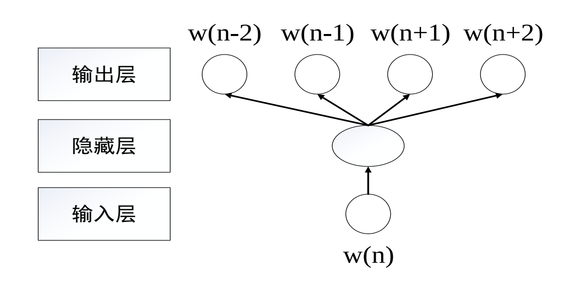


图 1　词向量训练模型结构示意图

假设存在一组词汇 $w_1,w_2,w_3,\dots,w_n$，Skip-gram 模型的目标是将下列公式最大化：

$$L=\frac{1}{N}\sum_{n=1}^{N}\sum_{-c\le i\le c,i\ne 0}\log p(w_{n+i}|w_n), \quad (1)$$

其中:c 是当前词语的前、后文词汇数，c 值越大则模型训练效果越优秀，但时间消耗也会增大。在实际应用中，若 c 选择恰当且训练语料库足够大，就能在短时间内得到高质量的词向量。在本文实验中，文本词汇被训练为 300 维的词向量，其训练网络是使用 Skip-gram 模型在包含 30 亿训练语料的 Google News 语料集上训练所得到的。之后，这些词向量被构建为文本矩阵并被用作 CNN 提取文本特征的输入数据。

2　CNN-ELM 混合短文本分类模型

如下图 2 所示是本文提出的 CNN-ELM 混合模型结构示意图，该模型使用经过词向量转换后的文本矩阵作为输入。卷积特征提取层使用不同大小的卷积核对输入矩阵进行特征提取，再对提取出的向量进行最大池化操作，之后进行拼接获得该文本矩阵的特征向量并使用多层 Highway 网络进行优化。最后，经过优化后的特征向量被当做极速学习机分类层的输入以完成最后的分类任务。

2.1　卷积特征提取层

从结构上来说，本文使用的卷积特征提取层是 Collobert 等人[21]所提出的 CNN 结构的一种变形。通过该层的处理之后，文本矩阵被转换为了特征向量。

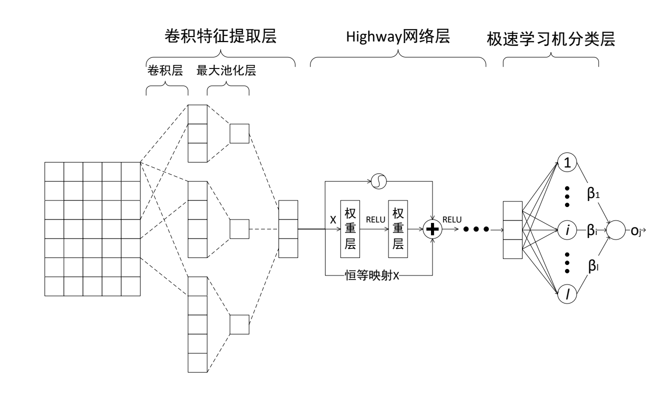


图 2　本文混合模型结构示意图

2.1.1 词向量的拼接

在输入数据的处理上，若 v_i 是句子中位置为 i 的 k 维词向量，n 是语料集中最长句子的长度，l 是该 CNN 中使用的卷积核的宽度最大值，那么，输入数据就是一个 $k*(n+l-1)$矩阵，这一个矩阵代表一个句子，它是由句子中所有词的词向量进行连接操作之后构成的，该连接操作可以表示为：

$$v_{1:n}=v_1\oplus v_2\oplus\dots\oplus v_n, \quad (2)$$

其中: $\oplus$ 是连接操作符。在本文模型中，为了更方便的处理数据，句子长度被设定为固定值（该数据集中的最大句长），使用填补操作对没有达到这一长度的句子矩阵填补$(l-1)$个 $\boldsymbol{0}$ 向量，最终得到的词向量个数是 $n+l-1$,他们的维度都是 k（=300）。

2.1.2 卷积操作

若卷积核的宽度 h，维度 k，其代表着该卷积窗口包含 h 个词向量，并利用这些词向量来产生一个新的特征。在本文模型中，卷积核宽度 h 是多样化的，通过结合不同的卷积窗口所提取出的特征向量可以更好地反映这一句子真正的语义特征。假定特征向量 a_i 是从词汇 $x_{i:i+h-1}$ 中产生的：

$$a_i=f(w\cdot x_{i:i+h-1}+b), \quad (3)$$

其中：w 是卷积核的权重，b 是偏置项，f 是非线性激活函数诸如 ReLU 或者 Tanh。类似于上述这种卷积核被应用在句子$\{x_{1:h},x_{2:h},\cdots,x_{n-h+1:n}\}$中从而生成一个特征向量 $\boldsymbol{a}$：

$$a=[a_1,a_2,\dots,a_{n-h+1}], \quad (4)$$

2.1.3 最大池化与特征向量输出

在获得了每个卷积核生成的特征向量之后，本文采取最大池化操作获取其最大值 $A=max\{a\}$，之后将 A 进行拼接，从而获得了卷积结构提取的特征向量。最大池化方法旨在获取不同卷积核所生成的最具代表性的特征，同时有效降低了空间与时间复杂度。

2.1.4 Dropout

为了避免训练过程中出现过拟合的情况，本文使用 Dropout 操作来禁止一部分隐层神经元参与前向传播，这使得这些神经元不参与此次更新过程，从而使权值的更新不依赖于固定节点

[1] https://code.google.com/p/word2vec/

的作用。之后，本文对 Dropout 操作后得到的向量使用 Highway 网络进行优化。

2.2 Highway 网络优化层

受 LSTM 的启发，Srivastava 等人提出的 Highway 网络也是一种可学习的门限机制，在该机制的作用下，Highway 网络能够对信息流进行局部调整从而实现信息流的优化。

在一个具有 L 层的传统前向神经网络中，每层网络都可对输入 x_i 使用具有参数 W_H 的非线性映射变换 H 产生输出 y_i，表示为

$$y = H(x, W_H), \tag{5}$$

Highway 网络在上述基础上增加了两个非线性映射函数 T 与 C，使得输出 y 变为

$$y = H(x, W_H)\cdot T(x, W_T) + x\cdot C(x, W_C), \tag{6}$$

其中:T 被称为转换门，C 被称为携带门。为了简化模型，携带门 C 通常被设置为$(1\text{-}T)$,则式(6)变为

$$y = H(x, W_H)\cdot T(x, W_T) + x\cdot(1-\mathrm{T}(x, W_C)), \tag{7}$$

其中:y 为 Highway 网络的最终输出。在式(7)中，x，y，H，T 的维度必须相同，若维度不足则进行补零操作。可以看到，Highway 网络通过转换门对其输入的信息流进行处理，这种类似于高速公路关卡的操作改变了部分输入信息流。这种方法已经被证实能够解决深度学习中模型难以训练收敛的问题从而提升模型性能。

在本文模型中，Highway 网络层实现了对卷积特征的优化，从而能够提高模型的分类效果。同时，由于 Highway 网络具有较低的复杂度，使用多层 Highway 网络连续优化卷积特征向量并不会增加过多的时间与空间损耗，本文通过实验确定 Highway 网络的最佳层数，从而获得最优特征向量。

2.3 极速学习机分类层

极速学习机自被提出之后便在分类与回归任务中展现出优异的性能。给定离散化的输入数据与隐层节点数，极速学习机能够快速计算出结果，这使得其在图像识别与手写数字识别中得到了广泛应用。对于 N 个不同的学习样本 $(\mathrm{x}_i, t_i)\in R_n\times R_m, (i=1,2,\ldots,N)$，极速学习机的基本原理可以被表示为：

$$\sum_{i=1}^{l}\beta_i g(w_i, b_i, x_j) = o_j,\ j=1,2,\cdots,N, \tag{8}$$

其中：$g(w_i, b_i, x_j)$ 表示极速学习机的隐层激活函数，β_i 是网络输出层和第 i 个隐层神经元的权值，b 为隐藏层神经元的偏置。上述公式可被简化为

$$H\beta = T, \tag{9}$$

其中:

$$\mathrm{H} = \begin{bmatrix} g(w_1,b_1,x_1) & g(w_2,b_2,x_1) & \cdots & g(w_l,b_l,x_1) \\ g(w_1,b_1,x_2) & g(w_2,b_2,x_2) & \cdots & g(w_l,b_l,x_2) \\ \vdots & \vdots & \ddots & \vdots \\ g(w_1,b_1,x_N) & g(w_2,b_2,x_N) & \cdots & g(w_l,b_l,x_N) \end{bmatrix}_{N\times l},$$

$$\beta = \begin{bmatrix} \beta_1^T \\ \beta_2^T \\ \vdots \\ \beta_l^T \end{bmatrix}_{l\times m}, T = \begin{bmatrix} t_1^T \\ t_2^T \\ \vdots \\ t_N^T \end{bmatrix}_{N\times m}。$$

此网络只需要设定隐藏层节点个数，不需要调整输入权值以及隐藏层上的偏置值，并且可以通过 $\boldsymbol{H}$ 矩阵和隐藏层上神经元的偏置值 b_j 求得网络隐藏层和输出层之间输出权值 $\beta: \beta = H^{\dagger}T$（其中：$H^{\dagger}$ 为 $\boldsymbol{H}$ 的 Moore-Penrose 广义逆矩阵）。

尽管极速学习机有着快速计算与泛化能力优异的特点，但想要计算出其最优解往往是通过人工统计的方式来进行的。误差最小化极速学习机(EM-ELM)[16]的提出解决了这一问题，它能够递归地向网络中添加隐层节点来自动计算出最佳结果。

给定训练集 $\{(x_i,t_i)\}_{i=1}^N$、最大隐层节点数 L_{max} 与期望学习准确率 $e>0$，EM-ELM 计算原理为如下两阶段：

阶段 1　初始化阶段

初始化单隐层前向神经网络（SLFN）并设置初始隐层节点数 $(a_i,b_i)_{i=1}^{L0}$（其中，L_0 为正整数，本文使用初始节点数 $L_0=1$）。之后和传统的 ELM 一样计算隐层输出矩阵 $\boldsymbol{H_1}$，并计算输出误差 $E(H_1)=\|H_1H_1^{\dagger}T-T\|$。

阶段 2　递归计算阶段

令 k=0.While $L_k<L_{max}$ and $E(H_k)>e$:

$k=k+1$,在已有的 SLFN 上随机添加 δL_{k-1} 个隐层节点，隐层节点数变为 $L_k = L_{k-1}+\delta L_{k-1}$，隐层输出矩阵变为 $H_{k+1}=[H_k, \delta H_k]$。之后使用递归方法更新输出权重 β：

$$D_k = ((I-H_kH_k^{\dagger})\delta H_k)^{\dagger}$$

$$U_k = H_k^{\dagger} - H_k^{\dagger}\delta H_k^T D_k$$

$$\beta_{k+1} = H_{k+1}^{\dagger}T = \begin{bmatrix} U_k \\ D_k \end{bmatrix} T$$

end while

在训练过程中，首先使用训练集将卷积神经网络层与 Highway 网络层训练至收敛，之后使用收敛后的模型对训练集与测试集文本矩阵进行特征提取，将提取后的训练集与测试集特征作为误差最小化极速学习机的输入，实现如图 2 的混合模型总体架构，该模型在误差最小化极速学习机进行初始化与递归计算后得到最佳分类结果。

3　实验与分析

本文使用多种英文短文本分类数据集来完成 CNN- ELM 混合短文本分类模型的性能测试。这些实验在拥有 256GB 内存与 Intel i7 4.0-GHz CPU 的服务器上使用 Python 2.7 与 TensorFLow 0.9.0 完成。

3.1　实验数据

为了评估本文模型，本文使用近些年被广泛应用于文本分类任务的英文短文本数据集来进行模型分类精确度的测试与分

析，这些数据集包含 MR[2]，SST-1[3]，Subj[22]，TREC[4]，Irony[23]，Tweet[5]，与 Polite[24]。同时，本文也爬取了社交媒体网站中的用户个人描述数据以构成 Description 数据集并进行了人工标注，经过一致性检验计算其 Kappa 系数为 0.83，满足数据集一致性需求，可以通过该数据集完成账户类型分类任务。上述各数据集的详细信息如下表 1 所示：

表 1　数据集详细信息

数据集	*c*	*l*	*N*	|*V*|	|V_{pre}|	*Test*
MR	2	20	10662	18765	16488	CV
SST-1	5	18	11855	17836	16262	2210
Subj	2	23	10000	21323	17913	CV
TREC	6	10	5952	9592	9125	500
Irony	2	75	1074	6138	5718	CV
Tweet	10	39	25552	33438	17023	5964
Polite	2	53	4353	10135	7951	CV
Description	4	68	5293	13709	8409	CV

其中：*c* 代表着目标类别数，*l* 代表着该数据集下的平均句子长度，*N* 代表着该数据集的容量，|*V*|代表着该数据集的词典容量，|*Vpre*|代表着该词典中在 Google 预训练词向量数据集中存在的词汇数，*Test* 是测试集容量。CV 代表着该数据集并没有进行训练集/测试集划分，所以本文使用十折交叉验证法（10-fold Cross Validation）来测试其分类性能。

3.2　超参数设置

3.2.1 词向量训练

本文使用 Google Word2Vec 工具进行词向量训练，每个单词被训练为一个 300 维的词向量，训练方法使用 Skip-gram。对于没有在 Worder2Vec 中出现的单词，本文对其进行随机初始化并赋予[-0.5，0.5]的随机值。

3.2.2 卷积神经网络

输入通道：CNN-non-static；

卷积核：本文使用宽度为[3，4，5]的卷积核各 100 个来进行卷积操作，激活函数为 ReLU；

Mini-Batch 规模：50；

Dropout 参数：0.5，仅在训练中有效；

优化器：ADAM 优化器；

学习率：0.001，每 8 次迭代衰减 50%；

3.2.3 Highway 网络

层数：2；

门偏置 bias：0；

非线性映射变换 H：ReLU；

3.2.4 误差最小化极速学习机

激活函数：Sigmoid；

最大隐层节点数 L_{max}：与数据集容量几乎相同，如表 2 所示。

表 2　数据集容量与最大隐层节点数设置

数据集	容量	*Lmax*
MR	10662	10000
SST-1	11855	10000
Subj	10000	10000
TREC	5952	6000
Irony	1074	1000
Tweet	25552	20000
Polite	4353	4000
Description	5293	5000

期望学习准确率 *e*：90%；

初始隐层节点数：初始为 1，逐次增加 1 个节点。

3.3　实验结果与对比分析

本文实验首先研究了 Highway 网络对分类结果带来的影响，通过对不同层数的 Highway 网络进行实验测试，确定出最佳 Highway 网络层数以构建本文分类模型。之后，将本文模型与多种机器学习算法相对比，其中既包含了多种传统机器学习算法及其改进算法（诸如 LDA、K 近邻、SVM、朴素贝叶斯、随机森林、决策树等），也包含了 Mandelbaum 等人改进的 CNN 文本分类模型。最后，研究了本文模型的不同网络结构所能对分类结果带来的影响。

3.3.1 Highway 网络对分类结果的影响

首先通过在 CNN 分类模型上结合不同层数的 Highway 网络来验证 Highway 网络对分类结果的影响，其在四种不同数据集上的分类结果如上图 3 所示。从图中可以看出， Highway 网络对不同数据集的优化能力并不同。在 Irony 与 Polite 数据集上，Highway 网络的对结果的提升最明显，在 Description 数据集上，Highway 网络也能够小幅提升分类结果。但是，在 TREC 数据集上是否添加 Highway 网络对结果几乎没有影响。

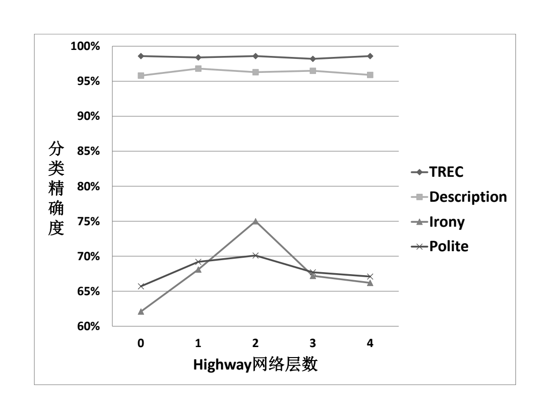


图 3　不同层数的分类结果对比图

就 Highway 网络层数来说，在 Irony 与 Polite 上的结果表明添加 2 层 Highway 网络进行特征优化时可得到最优结果，同

[2] https://www.cs.cornell.edu/people/pabo/moviereview-data/
[3] http://nlp.stanford.edu/sentiment/Data
[4] http://cogcomp.cs.illinois.edu/Data/QA/QC/
[5] http://www.cs.huji.ac.il/ nogazas/pages/projects.html

时在 Description 数据集上较原模型也有所提升。所以在之后的模型中本文都使用 2 层 Highway 网络进行特征优化操作。

从理论上来说，Highway 网络就是为了解决极深的神经网络难以训练的问题。几百甚至几千层的 Highway 网络可以直接使用梯度下降算法进行训练，并可以配合多种非线性激活函数，其优化方法基本与网络的深度独立，并不会过度增加时间与空间损耗。但是在实际训练过程中，由于训练数据量有限，过多的增加网络层数会导致网络过拟合从而使模型泛化能力下降，本文测试 100 层 Highway 网络在 Polite 数据集上的结果已低于 60%。但是，随着未来数据量的增加，极深 Highway 网络依旧有着非常大的潜能。

3.3.2 算法对比

表 3　与不同机器学习算法对比　/%

模型	精度
KNeighbors（k=4）	72.7
MultinomialNB	88.4
DecisionTree(entropy)	90.9
DecisionTree(gini)	91.5
RandomForest	93.1
Bagging	94.0
SVM(linear)	94.2
LDA	94.6
CNN-rand	95.7
CNN-static	93.8
CNN-non-static	95.8
本文模型	**97.3**

本文模型与传统机器学习模型在 Description 数据集上的实验结果如表 3 所示。可以看到，卷积神经网络的出现刷新了传统学机器学习算法的记录，其中，CNN-rand 与 CNN-non-static 结果相近，这是因为在网络数据中存在大量的新词、表情、链接地址从而导致该数据集的词典中将近一半的词都没有在 Word2Vec 中出现过，这也使得随机赋值的词向量在词典中比重较大，使用随机模型 CNN-rand 也就与词向量模型 CNN-non-static 相差甚微。在卷积神经网络的基础上，本文模型首先使用 Highway 网络对卷积神经网络提取的特征进行优化，在 CNN-non-static 基础上提高了 1%的分类精度，同时结合极速学习机分类原理，最终将分类结果提高到了 97.3%。

表 4　不同数据集上的结果对比　/%

模型	MR	SST-1	Subj	TREC	Irony	Tweet	Polite
CNN-non-static	80.5	47.3	93	98.6	62.1	**89.2**	65.7
本文模型	**81.4**	**51.8**	**94.3**	**99.4**	**76.6**	88.1	**71.3**

本文使用多种短文本分类数据集进一步测试本文模型的泛化能力，其与改进的 CNN-non-static 模型结果对比如表 4 所示。可以看出，本文模型在多数数据集上获得了更高的分类精度。在 TREC 数据集上，图 3 已表明 Highway 网络无法有效提升分类效果，但本文模型最终使结果提高了 0.8%，这也进一步证明了结合极速学习机分类原理的有效性。同时，因为 2 层 Highway 网络在 Irony 与 Polite 数据集上有着显著的优化能力，在没有与极速学习机结合时就分别能达到 75%与 70.1%的精度，所以与极速学习机相结合就能进一步提高模型分类性能。

3.3.3 本文模型的网络结构对分类结果的影响

表 5　本文模型不同结构的结果对比　/%

模型结构	Description	SST-1	Subj	Tweet	Polite
CNN-non-static	95.8	47.3	93.0	89.2	65.7
CNN-ELM	96.1	48.6	93.3	88.2	68.1
CNN-1Highway	96.8	48.5	93.5	**89.6**	69.2
CNN-2Highway	96.3	49.8	93.7	88.2	70.1
CNN-1Highway-ELM	**97.3**	48.8	93.7	88.1	70.2
CNN-2Highway-ELM	96.7	**51.8**	**94.3**	88.1	**71.3**

本文模型在 CNN 基础上结合 Highway 网络与极速学习机原理以完成分类任务，这三种结构的不同结合形式及其在不同数据集上的实验结果对比如上表 5 所示。可以看出，大多数情况下仅结合 Highway 网络比仅结合极速学习机对混合模型的提升更大，在 Tweet 数据集上仅结合 1 层 Highway 网络的模型更是获得了最高精确度。通常来讲，在 Highway 网络优化后结合极速学习机往往能获得最佳结果。同时，在 Description 数据集上的实验表明选择正确的 Highway 网络层数对最佳结果也会产生较大影响。综上所述，若想获得特定任务的最优模型依旧需要对参数与模型进行多次调整。然而，通过人工测试确定最佳模型的方法过于复杂，若能够使 Highway 网络自动选择最佳层数完成分类任务，该混合模型的可靠性就能大幅提高。

4　结束语

本文提出了一种 CNN-ELM 混合短文本分类模型，在卷积神经网络的基础上，该模型结合了极速学习机与 Highway 网络的相关理论，获得了较原有模型更优秀的分类结果。实验表明，该方法比传统机器学习算法与卷积神经网络模型更有效。未来研究工作包括以下几方面：改进 Highway 网络，通过增加能够自动选择最佳网络层数的功能，增强混合模型的可靠性；通过添加外部特征并与卷积特征相结合以提高分类性能；对卷积神经网络特征提取结构进行改进，通过与其他深度神经网络相结合构成更深层次的网络结构以提取出更具代表性的特征。

参考文献：

[1] Cover T M, Thomas J A. Elements of information theory [M]. 2nd ed. New Jersey: Brooks//John Wiley & Sons, 2012.

[2] Cai Lijuan, Hofmann T. Text categorization by boosting automatically extracted concepts [C]// Proc of the 26th Annual International ACM SIGIR Conference on Research and Development in Informaion Retrieval. 2003:

182-189.

[3] Hingmire S, Chougule S, Palshikar G K, et al. Document classification by topic labeling [C]// Proc of the 36th International ACM SIGIR Conference on Research and Development in Information Retrieval. 2013: 877-880.

[4] Bengio Y, Ducharme R, Vincent P, et al. A neural probabilistic language model [J]. Journal of machine learning research, 2003, 3 (Feb): 1137-1155.

[5] Mikolov T, Sutskever I, Chen Kai, et al. Distributed representations of words and phrases and their compositionality [C]// Neural Information Processing Systems. 2013: 3111-3119.

[6] Kim Y. Convolutional neural networks for sentence classification [C]// Proc of Conference on Empirical Methods in Natural Language Processing. 2014: 1746-1751.

[7] Mandelbaum A, Shalev A. Word embeddings and their use in sentence classification tasks [J/OL]. arXiv preprint, 2016, arXiv: 1610. 08229 [2017-11-30]. https://arxiv. org/abs/1610. 08229.

[8] 陈钊，徐睿峰，桂林，等. 结合卷积神经网络和词语情感序列特征的中文情感分析 [J] 中文信息学报, 2015, 29 (6): 172-178.

[9] 刘龙飞，杨亮，张绍武，等. 基于卷积神经网络的微博情感倾向性分析 [J] 中文信息学报, 2015, 29 (6): 159-165.

[10] 肜博辉，付琨，黄宇，等. 基于多通道卷积神经网的实体关系抽取 [J]. 计算机应用研究, 2017, 34 (3): 689-692.

[11] Ebert S, Vu N T, Schütze H. CIS-positive: combining convolutional neural networks and svms for sentiment analysis in Twitter [C]// Proc of the 9th International Workshop on Semantic Evaluation. 2015: 527-532.

[12] Matsugu M, Mori K, Suzuki T. Face recognition using SVM combined with CNN for face detection [C]// Proc of International Conference on Neural Information Processing. 2004: 356-361.

[13] Huang Guangbin, Zhou Hongming, Ding Xiaojian, et al. Extreme learning machine for regression and multiclass classification [J]. IEEE Trans on Systems, Man, and Cybernetics, Part B: Cybernetics, 2012, 42 (2): 513-529.

[14] Huang Guangbin, Zhu Qinyu, Siew C K. Extreme learning machine: a new learning scheme of feedforward neural networks [C]// Proc of IEEE International Joint Conference on Neural Networks. 2004: 985-990.

[15] Yu Jiasheng, Chen Jin, Xiang Z Q, et al. A hybrid convolutional neural networks with extreme learning machine for WCE image classification [C]// Proc of IEEE Conference on Robotics and Biomimetics. 2015: 1822-1827.

[16] Feng Guorui, Huang Guangbin, Lin Qingping, et al. Error minimized extreme learning machine with growth of hidden nodes and incremental learning [J]. IEEE Trans on Neural Networks, 2009, 20 (8): 1352-1357.

[17] Srivastava R K, Greff K, Schmidhuber J. Highway networks [J/OL]. arXiv preprint, 2015, arXiv: 1505. 00387 [2017-11-30]. https://arxiv. org/abs/1505. 00387.

[18] Srivastava R K, Greff K, Schmidhuber J. Training very deep networks [C]// Advances in Neural Information Processing Systems. 2015: 2377-2385.

[19] Kim Y, Jernite Y, Sontag D, et al. Character-Aware Neural Language Models [C]// Proc of the 30th AAAI Conference on Artificial Intelligence. 2016: 2741-2749.

[20] Hsu W N, Zhang Y, Lee A, et al. Exploiting depth and highway connections in convolutional recurrent deep neural networks for speech recognition [C]// Proc of InterSPEECH. 2016: 395-399.

[21] Collobert R, Weston J, Bottou L, et al. Natural language processing (almost) from scratch [J]. Journal of Machine Learning Research, 2011, 12 (Aug): 2493-2537.

[22] Pang Bo, Lee L. A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts [C]// Proc of the 42nd Annual Meeting on Association for Computational Linguistics. 2004: 271.

[23] Wallace B C, Do Kook Choe L K, et al. Humans require context to infer ironic intent (so computers probably do, too) [C]// Procof the 52nd Annual Meeting of the Association for Computational Linguistics) . 2014: 512-516.

[24] Danescu-Niculescu-Mizil C, Sudhof M, Jurafsky D, et al. A computational approach to politeness with application to social factors [J/OL]. arXiv preprint, 2013, arXiv: 1306. 6078 [2017-11-30]. https://arxiv. org/abs/1306. 6078